Ready. Set. 2.0!

The NEW SunCHECK™

INDEPENDENT QA.
YOUR WAY.

"Having patient information, having machine information all in one easily accessible place... ***that's basically what you would want in a system.***"

Robert Staton, Ph.D., UF Health Cancer Center at Orlando Health, Orlando, FL

Learn more at:
sunnuclear.com

Platform

Patient

Machine

SUN NUCLEAR corporation

A neural network approach for fast, automated quantification of DIR performance

John Neylon[a)], Yugang Min, Daniel A. Low, and Anand Santhanam
Department of Radiation Oncology, UCLA, 200 Medical Plaza, Suite B265, Los Angeles, CA 90095, USA

(Received 31 August 2016; revised 13 April 2017; accepted for publication 30 April 2017; published 17 July 2017)

Purpose: A critical step in adaptive radiotherapy (ART) workflow is deformably registering the simulation CT with the daily or weekly volumetric imaging. Quantifying the deformable image registration accuracy under these circumstances is a complex task due to the lack of known ground-truth landmark correspondences between the source data and target data. Generating landmarks manually (using experts) is time-consuming, and limited by image quality and observer variability. While image similarity metrics (ISM) may be used as an alternative approach to quantify the registration error, there is a need to characterize the ISM values by developing a nonlinear cost function and translate them to physical distance measures in order to enable fast, quantitative comparison of registration performance.

Methods: In this paper, we present a proof-of-concept methodology for automated quantification of DIR performance. A nonlinear cost function was developed as a combination of ISM values and governed by the following two expectations for an accurate registration: (a) the deformed data obtained from transforming the simulation CT data with the deformation vector field (DVF) should match the target image data with near perfect similarity, and (b) the similarity between the simulation CT and deformed data should match the similarity between the simulation CT and the target image data. A deep neural network (DNN) was developed that translated the cost function values to actual physical distance measure. To train the neural network, patient-specific biomechanical models of the head-and-neck anatomy were employed. The biomechanical model anatomy was systematically deformed to represent changes in patient posture and physiological regression. Volumetric source and target images with known ground-truth deformations vector fields were then generated, representing the daily or weekly imaging data. Annotated data was then fed through a supervised machine learning process, iteratively optimizing a nonlinear model able to predict the target registration error (TRE) for given ISM values. The cost function for sub-volumes enclosing critical radiotherapy structures in the head-and-neck region were computed and compared with the ground truth TRE values.

Results: When examining different combinations of registration parameters for a single DIR, the neural network was able to quantify DIR error to within a single voxel for 95% of the sub-volumes examined. In addition, correlations between the neural network predicted error and the ground-truth TRE for the Planning Target Volume and the parotid contours were consistently observed to be > 0.9. For variations in posture and tumor regression for 10 different patients, patient-specific neural networks predicted the TRE to within a single voxel $> 90\%$ on average.

Conclusions: The formulation presented in this paper demonstrates the ability for fast, accurate quantification of registration performance. DNN provided the necessary level of abstraction to estimate a quantified TRE from the ISM expectations described above, when sufficiently trained on annotated data. In addition, biomechanical models facilitated the DNN with the required variations in the patient posture and physiological regression. With further development and validation on clinical patient data, such networks have potential impact in patient and site-specific optimization, and stream-lining clinical registration validation. [https://doi.org/10.1002/mp.12321]

Key words: accuracy, DIR, machine learning, neural network, radiotherapy

1. INTRODUCTION

Adaptive radiation therapy (ART) has potential for improving the efficacy of cancer therapy by adapting the plan to a patient's daily anatomy. Several recent studies have shown that ART can provide significant dosimetric benefits for inter-fraction anatomic variations, as well as reduced normal tissue toxicity, in the head-and-neck,[1–4] as well as other cancer sites.[5–8] ART may also allow a reduction in the error margins added around the clinical tumor volume (CTV) to construct the planning target volume (PTV).[9] To institute online ART, previously delivered dose must be accumulated and mapped to the daily anatomy while the patient is on the treatment couch, greatly shortening the time scales for registration and validation.[10] This increased manpower requirements have inhibited full online capabilities for daily monitoring of every

patient.[10] Clinical implementations of ART remain largely limited to offline studies and require a significant amount of user intervention.[1,11] Methods and tools to enable an online quantification of the ART's reliability must be developed before implementing online ART in a clinical setup, and forms the focus of our research.

Deformable image registration (DIR) allows image-guided analyses of nonrigid anatomical variations.[12,13] DIR has multiple applications in ART,[14] including dose accumulation and contour propagation.[15] Computational speed and accuracy of DIR is critical to quantitatively track changes in patient anatomy and to the overall success of plan adaptation. There remains an unmet need for an automated quantification of DIR accuracy that is tailor-made for the anatomy and image modality involved. There has been much work in recent years assessing and comparing the accuracy of commercially available DIR algorithms.[16–20] However, these studies have limited applicability for online ART as the DIR accuracy needs to be quantified on a per-patient basis.[21] The current gold standard for obtaining clinical registration error requires comparison between manually placed corresponding landmarks on the source and target and calculating the difference between the user-defined deformation and the deformation reported by the DIR,[22] commonly referred to as target registration error or TRE. However, placing landmarks is time intensive, subject to inter- and intra-observer variability, and suffers from small sample size.[23,24] For clinical scenarios, since the true deformation remains unknown, automating the landmark assessment process is not possible. Clinical DIR assessment has also been hampered by lack of techniques to generate ground-truth deformations that represent patient geometry changes (posture and physiological regression) for evaluating and quantifying DIR performance. Due to these hindrances, formulations to quantitatively assess the accuracy of online clinical registrations are not readily available.[25]

Image similarity metrics (ISMs) provide a fast method for assessing the correlation between two image sets and outputs a single value quantifying the similarity between intensity fields. However, the quantification has little physical meaning without a proper frame of reference.[26] Therefore, using image-based metrics is currently qualitative, that is, the range of values for each ISM does not correspond directly to a physical error measure. In 2012, Rohlfing presented an exhaustive study of the limitations of image similarity and tissue overlaps as accuracy measures for deformable image registration.[27] He showed how direct application of these measures can be deceptive, reaffirming the lack of an automated method for quantifying DIR accuracy. It was concluded that the gold standard remains manually placed landmarks, despite the low efficiency and time requirement of the method.

A fast automated methodology for assessing registration performance is necessary for implementation into the daily clinical workflow.[10] In addition, there is also an opportunity to improve DIR accuracy by optimizing registration parameters on a per-patient or per-registration basis[28] and site-specific basis,[29,30] demonstrating the need for narrow focus of DIR to the current clinical application. Kirby et al. presented an automated tool for evaluating DIR algorithms in general,[31] but patient-specific assessment of clinical registrations remains elusive. A fast, automated methodology for assessing registration performance is necessary for per-patient or site-specific registration optimization to be implemented into a daily clinical workflow. Towards these goals, we first present a nonlinear cost function using a combination of ISMs. The ISM computations at this stage are fully automated and near real-time. We then present a machine-learning approach using deep neural networks (DNN).

The potential of a neural networks as quality evaluators for rigid transformations during head-and-neck patient set up has been shown previously,[32,33] but such an effort has not been investigated for deformable image registrations, which forms the key contribution of this paper. Neural networks have gained significant traction in recent years in a wide variety of fields. The advantage of deep neural networks lies in their ability to learn relationships from annotated data, without the necessity for user intervention to design specific features or identifiers. This is typically done using a form of stochastic gradient descent to modify weights and biases until the network output matches the expected results as closely as possible. The depth and scope of the network allows it to construct complex relationships, and once trained, to accurately infer a result from unlabeled input data. However, application of neural networks in the medical arena have been predominantly focused on the field of computer aided diagnosis.[34–39]

Our DNN approach involved creating large correlated data sets of model-generated synthetic CTs representing the simulation CT and the subsequent daily or weekly patient imaging. The network was then able to translate the ISM expectations into average TRE values. This work proves that a neural network approach can accurately quantify DIR performance from image similarity values for model-generated deformations, and provides a foundation for building a neural network to quickly and accurately assess clinical registrations.

2. METHODS

2.A. Expectations of the ISM response

The first step towards our automated registration accuracy quantification process was to develop a nonlinear cost function which utilizes the image similarity metrics. To this end, an initial cost function was proposed based on the expectation that for a good registration, in addition to having a high similarity between the deformed and target images, the deformed image will be similar to the target image when both are processed in comparison to the source image. This provides a frame of reference to assess the result relative to the initial similarity of the source and target. Registrations are inherently image specific, and image similarity does not provide a one-to-one correspondence to registration quality. The same similarity may represent a poor registration for two images that already have high similarity, and the best possible registration between two images that are very different. The

additional expectation term, comparing both deformed image and target image to the source image, provides a method of standardization to some degree, and allows comparisons between registrations of varying difficulties.

In our approach, the cost function was developed around two expectations as the registration error approaches zero. The given image pairs are considered to be the source and target images. The warped dataset was created by applying the deformation vector field (DVF) obtained from the DIR algorithm to the source image. Similarity measures were calculated for three sets of images: source-target (I_{ST}), source-warp (I_{SW}), and target-warp (I_{TW}). The expectations can then be expressed as: [Eq. (1)] the ISM value representing the similarity between the target and warped (or deformed) datasets should approach 1, and [Eq. (2)] the similarity between the source and warped datasets should approach the similarity between the source and target datasets.

$$Y = I_{TW} \tag{1}$$

$$X = 1 - |I_{ST} - I_{SW}| \tag{2}$$

For an ideal registration, $X \to 1$, and $Y \to 1$. The image similarity metric chosen for testing initial response was normalized mutual information (NMI), which uses the Shannon entropy of the individual images sets, H_A and H_B, and their combined entropy, H_{AB}. Here the entropy was calculated from the histogram of image intensities, where $p(x)$ represents the probability of each histogram bin. The expressions for the normalized mutual information and Shannon entropy are shown in Eq. 3(a) and 3(b), respectively.[40,41]

$$NMI = \left(\frac{H_A + H_B}{H_{AB}}\right) - 1, \tag{3a}$$

$$with\ H_A = -\sum\nolimits_A p(a)\, log\, p\ (a) \tag{3b}$$

Equation 4 shows the proposed cost function combining the similarity terms from Eq. (1) and Eq. (2), where m and n are variables to be optimized, and f is a weighting factor between 0 and 1. A systematic analysis was performed to determine the effect of the cost function variables (CFVs) (m,n,f) on the cost function response (CFR). An inverse near-linear relationship was expected between the ISM values and the average ground truth TRE (gt-TRE), which is defined as the Euclidean distance between the deformation vector calculated by the DIR algorithm and the true deformation vector. By adjusting the CFVs, the response curve can be manipulated.

$$CFR = fX^m - (1-f)Y^n \tag{4}$$

2.B. Establishing a predictive relationship between the ISM cost function and TRE

Since the correlation between the ISM and the TRE is not explicit, it can be construed that a direct correlation might not exist for the ISM cost function and TRE. To account for this, we developed a deep neural network model to transform the ISM expectation values described by Eq. (1) and Eq. (2) directly to the average TRE values. The inclusion of the second expectation provides additional value as another input neuron to the neural network, allowing more connections to be established and increasing the complexity of the characterization.

2.B.1. Neural network construction

The neural network employed in this work is a fully connected three-layer network. As inputs, the values from Eq. (1) and (2) were calculated for the sub-volumes encompassing four critical structures in head-and-neck radiotherapy: the primary PTV, left parotid, right parotid, and cord. The output of the network was a vector of neural network predicted average TRE (nn-TRE) values corresponding to the volumes encompassing each of these structures. The number of neurons in the hidden layer was optimized for the best result, ultimately settling at thirteen. A simple schematic of the network is shown in Fig. 1.

Annotated data was split between a training set and a test set as 25% to 75%, respectively. As the training data (further discussed in section 2.C) was fed through the network, a series of weights and biases were optimized to minimize a loss function. The accuracy of the network was continually monitored by inferring an output from the test data, and comparing to the ground truth expectations. Figure 2 illustrates the flow of data for the full network architecture. The eight input values were sent through the hidden and output layers, while the four known expectations are sent to the loss and accuracy functions. The result of the output layer was then sent to the loss function, accuracy function, and training algorithm, which updated the weights and biases of the hidden and

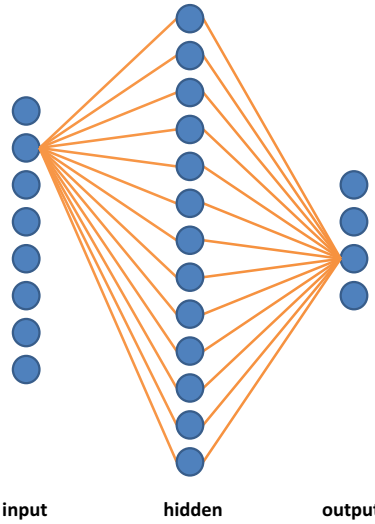

FIG. 1. A conceptual representation of a three-layer fully connected network is shown with 8 input neurons, 13 hidden neurons, and 4 output neurons. [Color figure can be viewed at wileyonlinelibrary.com]

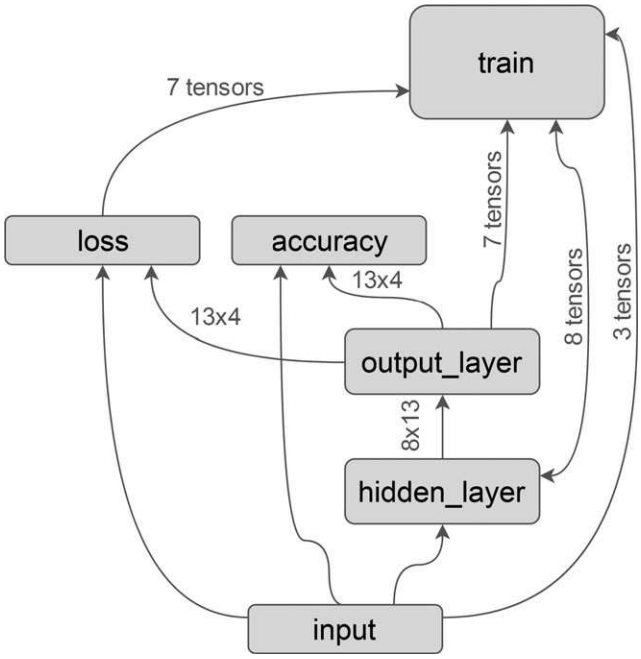

FIG. 2. A graph visualization of data flow for the network used in this manuscript is shown, constructed using the TensorBoard graph visualization tool provided in the TensorFlow library.

output layers and were used to compute the network accuracy (further discussed in §2.B.3).

2.B.2. Neuron activation function

The neurons of the hidden layer took the data from each neuron of the input layer, apply a matrix multiplication with weighting factors, add a bias, and then apply a nonlinear activation function. Without this activation function, the network would be comprised of a combination of linear functions, shown in Eq. (5), where the activation of a hidden neuron, a_j, is a linear function defined by the input, z, weight, w, and bias, b, summed over the input neurons, i. These weights and biases are the values adjusted during training and allows the network to learn.

$$a_j = \sum_i w_{j,i} z_i + b_j \tag{5}$$

Converting this to a nonlinear response is important because a composition of linear functions remains a linear function, so the network abstraction is limited no matter its depth. The activation function chosen for this network was the sigmoid function [Eq. (6)].[42]

$$\sigma(a) = \frac{1}{1 + e^{-a}} \tag{6}$$

The sigmoid function is essentially a smoothed out step function. The sigmoid function was chosen because there was no loss of data for negative values, which is typical for other activation functions, such as hyperbolic tangent and rectified linear unit function. This was important because the network outputs a physical value. Sigmoid activation has the drawback of possibly saturating during training, but this was not much of a concern for the size of the network being employed in this manuscript.

2.B.3. Loss and accuracy measures

The loss function was applied during training to calculate the error between the output of the feed-forward neural network TRE inference and the gt-TRE. Since the output of the network was intended to be a physical quantity, quadratic cost was implemented as the loss function.[42,43]

$$Loss = \frac{1}{N_s} \sum_s (\boldsymbol{y_s} - \boldsymbol{a_s})^2 \tag{7}$$

Where N_s is the size of the training data set, s is the individual training data, and $\boldsymbol{y}$ was the tensor of true expected outcomes, and $\boldsymbol{a}$ was the tensor of network outputs.

The accuracy was calculated [in Eq. (8)] as an absolute percent error between the nn-TRE and the gt-TRE, with a target of 0.1 mm accuracy. Relative percent error had little physical meaning for instances where the gt-TRE approached zero. The denominator of 2 [in Eq. (8)] corresponded to an expected value of 2 mm for the gt-TRE, y. Therefore, 75% accuracy corresponded with a physical margin of 0.5 mm, and an accuracy of 0% corresponded with an error of 2 mm from the actual gt-TRE. Setting the expectation value at 2 mm matched the in-plane resolution of the CT data being registered, which had voxel dimensions of 1.953 mm by 1.953 mm with a slice thickness of 3 mm. Using this measure, any value between 0 and 100% can be considered sub-voxel accuracy, approaching 0 error at 100% accuracy, and any errors greater than 2 mm would extend into the negative region.

$$Accuracy = 100 * \left(1 - \frac{|y_s - a_s|}{2}\right) \tag{8}$$

2.B.4. Back-propagation

In order for the network to be trained, the error from the loss function has to alter the network to better approximate the expected outcome. Backpropagation is a method of retracing the network from output to input, adjust weights, w, and biases, b, at each layer by applying their respective partial derivatives of the loss function.[44] Rumelhart et al. showed the performance benefits of backpropagation utilizing gradient descent.[45] Currently, the most widely used approach for neural network training is stochastic gradient descent.[46] Stochastic gradient descent (SGD) trains on smaller batches of training data, called epochs. Within an epoch, the training batch is iterated through several times, randomly choosing data points to estimate the gradient. An adaptive sub-gradient method, with dynamic learning rates was employed for network training.[47]

2.C. Generating ground truth data

Training data is critical in enabling the deep neural network to precisely predict the nn-TRE for given values of the ISM. The key requirement for such a training data is to have a large number of voxels with known ground truth TRE values. Generating such a large data using clinical datasets is a tedious task as most manual DIR validations consider 300 landmarks or more. To address this issue, we employed a biomechanical model to generate automated ground truth TRE values as further described below.

2.C.1. Simulated CTs with known DVFs from biomechanical modeling

In previous work, a framework was developed with the ability to instantiate an interactive biomechanical model from a patient CT.[48] These models were used to induce posture changes and physiological regression to simulate day-to-day changes in patient anatomy and create clinically realistic ground truth deformation vector fields for the purpose of clinical DIR validation. The model was validated to reproduce clinically observed deformations, including posture changes and tumor regression, with a correlation coefficient greater than 0.9. Figure 3 illustrates through volume renderings the systematically induced deformations achieved using the biomechanical model, including tumor regression and rotations of the head. These deformations reasonably simulated clinically observed changes to patient anatomy and provide a reasonable testing dataset for the automated DIR assessment methodology. The framework outputs a simulated CT of the deformed anatomy and a fully volumetric DVF so the motion of each voxel was known.

2.C.2. Dense registration parameter space/TRE correspondence

An in-house multi-level, multi-resolution optical flow DIR algorithm was employed for these experiments.[49] The registration algorithm had four adjustable parameters: (a) the smoothing factor, (b) the number of resolution levels, (c) the number of iterations, and (d) the number warps. Registrations were performed for a systematic sampling of this four-dimensional registration parameter space. Table I displays the sampling rate for each variable and the total number of registration performed. A total of 2400 DIR computations were performed between the source-target dataset to create the dense parameter space data set. The induced deformation of the target image for this data set consisted of 15° rotations about each axis, and 25% regression in the primary tumor contour. This was a much larger change in anatomy than typically observed clinically, but was chosen to accentuate the differences For each registration, a deformed image volume was created from the DIR DVF, the gt-TRE calculated, and similarity analysis was run between the three sets of image pairs.

Additionally, annotated data was generated for a variety of anatomies by inducing 45 different postures with the biomechanical model, systematically rotating the head about the three primary axes. At each posture, six levels of regression were applied to the primary tumor target, creating a total of 270 target volumes with known deformations. Registrations were run between the source and each of these target volumes for five different smoothing parameters, and the gt-TRE was recorded by randomly and automatically selecting 100 landmarks within each structure of interest and comparing the DIR DVF with the known model DVF. Table II describes the composition of this multi-pose anatomy data set.

2.C.3. Sub-volume/site specific assessments

Analyzing the similarity of CT images of the head-and-neck at a full volumetric level can diminish the effectiveness due to the high percentage of empty space, and the lack of deformation in areas such as the brain. Therefore, analysis was also performed on sub-volumes of the data. These sub-volumes were automatically generated with respect to the extents of the contoured structures of interest for radiotherapy purposes, including the right and left parotid glands, and the tumor targets.

2.D. Development environment

The biomechanical model, registration algorithm, and image similarity analysis tools were developed in a Linux environment, using C/C++ and accelerated with NVIDIA's CUDA library to run on graphics processing units (GPUs). Neural network development was done in python, using the Google's open source library for machine intelligence, TensorFlow, accelerated for GPU with the CUDA deep neural network library, cuDNN.

3. RESULTS

In this section, we first present the correlation between the ISM cost function and the ground truth TRE. This illustrates the lack of an explicit correlation between the two metrics, supporting the need for a deep learning based approach. We then discuss the results obtained from the DNN correlation between the ISM cost function and the ground truth TRE. As part of this effort, we first present the results for variations in the DIR parameters. It is followed by a discussion on the DNN accuracy for variations in the patient posture and physiological changes.

3.A. Cost function response versus target registration error

Figure 4 shows the full 4D parameter space stretched over the *x*-axis, with plots of the gt-TRE and CFR in the primary and secondary y-axes, respectively, for the PTV1 volume. Within each level subdivision in the figure, there are 8 peaks corresponding to the varying number of iterations. The

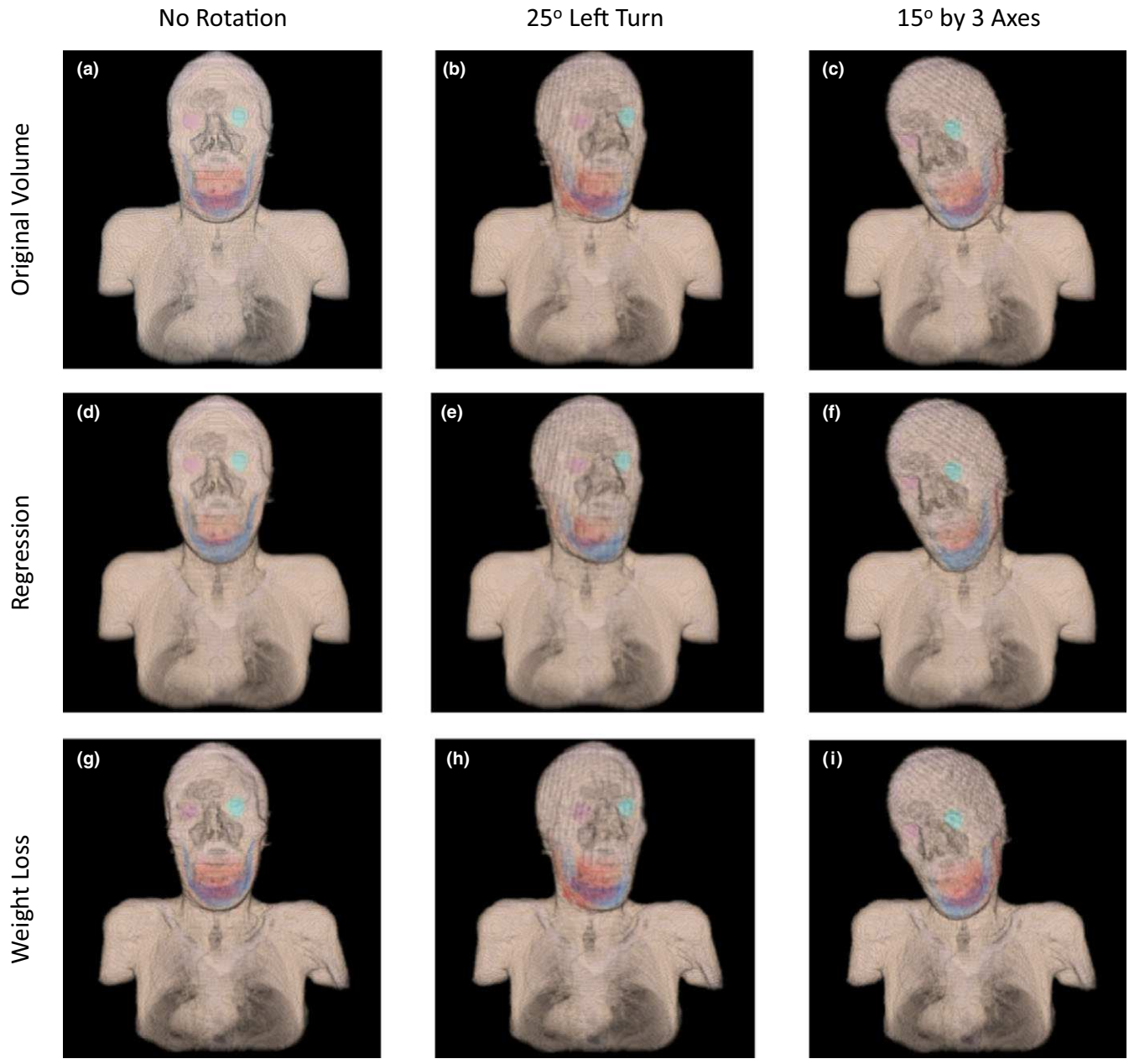

FIG. 3. Renderings of a patient-specific biomechanical model for a variety of posture and volume change combinations. [Color figure can be viewed at wileyonlinelibrary.com]

smooth curves between peaks correspond to the sampling of the smoothing parameter. As expected, there was an observable inverse correlation. While it appears that the moving average of the CFR increased in concordance with the moving average of the gt-TRE decreasing, the actual correlation between the data was just −0.4675 for the cost function variables used to generate the data in Fig. 4. When plotted against each other, shown in Fig. 5, we observed that the cost function was able to establish a general trend of decreasing response for larger TRE values, but the lack of a one-to-one correspondence suggest the cost function was not sophisticated enough to capture the full relationship between TRE and image similarity metrics.

To investigate the CFR correlation, a systematic evaluation of the cost function variables (CFV) was performed. Our results show that the CFR showed strong dependence on the CFV. For the data in Figs. 4 and 5, the weighting factor, f, was set to 0.5, and the exponents, m and n, were set as 2 and 0.5, respectively. This CFV set will be referred to as the reference CFV from this point on. Figure 6 shows how the CFR varied for different sets of CFV. The data presented comes from the sub-volume surrounding the right parotid gland. For the right parotid, a high correlation was found (−0.92) by adjusting the CFV, but resulted in poor correlation for the other sites being examined. We found that no matter how the CFV were modified, a consistently good correlation across all sites with a

single CFV set was not achieved. This reinforced the assertion that the relationship was too complex for the proposed cost function, and motivated our investigation of a neural network approach, the results of which are now discussed.

3.B. Neural network results

The results of the cost function experiments indicated that a more complex representation was necessary to fully characterize the relationship between the ISM expectation terms, *X* and *Y*, as described by Eq. 1 and 2. A neural network was developed which took as inputs these ISM expectation terms and would infer a target registration error (nn-TRE) as an output. In this section, the results for the neural network are presented, where manually adjusted variables are eliminated and replaced with a framework for learning from annotated training data.

TABLE I. Sampling frequency for each parameter of the 4D registration parameter space. A total of 2400 registrations were performed between a single source-target image set to create the dense parameter space data set.

Registration parameter	Range	Instances
Warps	1:3	3
Levels	1:5	5
Iterations	50:500	8
Smoothing	10:1000	20
Total registrations		2400

3.B.1. Training on the dense parameter space data set

The first experiment with the neural network re-examined the dense parameter space data set, which consisted of 2400 registrations between a single sets of volumetric images. By systematically varying the registration parameters, subtle differences were induced in the DVF, spanning the quality spectrum of registrations. The details of this experiment was discussed in §2.C.2.

From the 2400 samples in the dense parameter space data set, 25% (600) were chosen randomly as the training data.

TABLE II. Composition of annotated training data for systematic variations in head posture and tumor regression levels. Additionally, the smoothing parameter of the DIR algorithm was varied to create a total of 1350 registrations for the multi-pose anatomy data set.

	Levels of regression	Postures			Registration smoothing	Annotated data sets
Instances	6	45			5	1350
Range	0:30%	x-rotation	y-rotation	z-rotation	50:1000	
		−4:4	−2:2	−2:2		

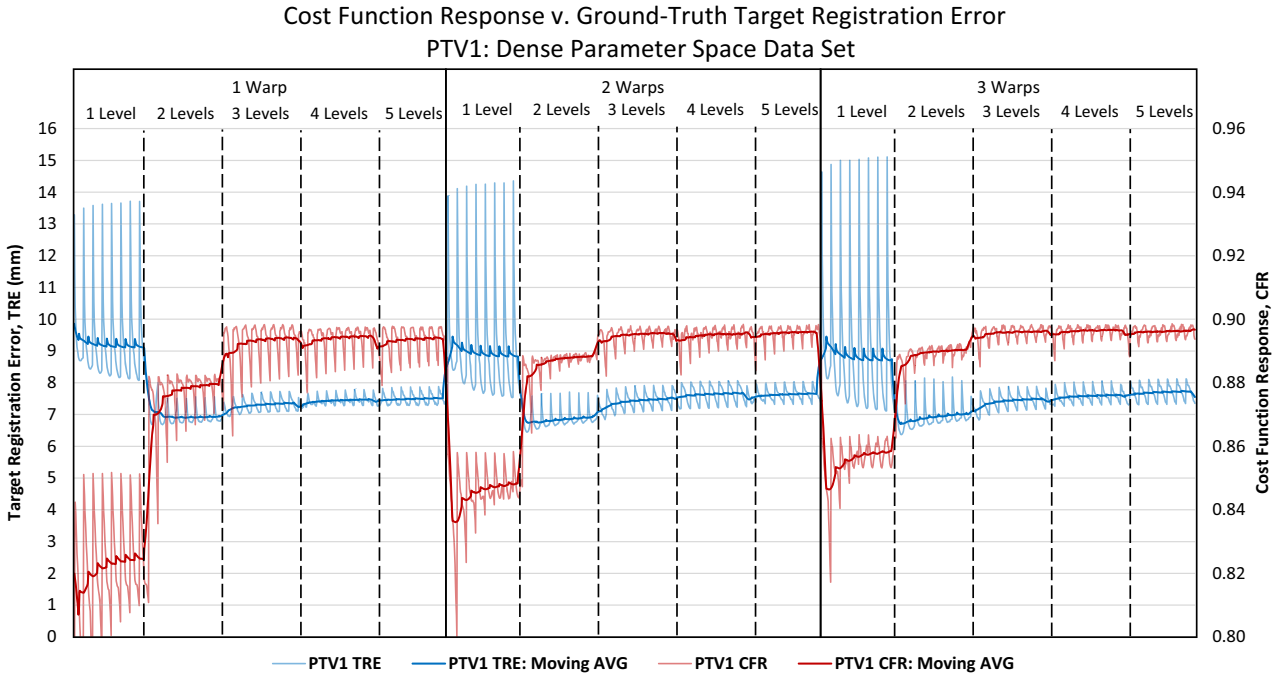

FIG. 4. Comparison of TRE and CFR over the entire dense parameter space data set for the PTV1 contour, with moving averages for a window size of 20 samples. Four registration parameters were systematically sampled. Number of warps comprised the outermost loop, followed by number of levels, iterations, and smoothing value. The plot shows delineations of how the 4D parameter space was plotted in 1D along the *x*-axis for the warps and levels. An inverse correlation can be observed between TRE and CFR throughout the entire registration parameter space. [Color figure can be viewed at wileyonlinelibrary.com]

The network was trained in batches of 75 samples over 1000 epochs, such that every 8 epochs, all 600 training data had been iterated through. After training, the entire data set was fed through the network. The network reached 88% accuracy on the test data, which consisted of 75% of the dense parameter space data set. The results are shown in Fig. 7, along with the difference in millimeters between the predicted TRE and the true TRE. The network was able to predict the TRE to within 1 mm for the entire registration parameter space, excluding the purposely poor registrations performed using only 1 warp, 1 level, and less than 100 iterations. The mean accuracy for the PTV1 was just under 86%, corresponding to a mean discrepancy less than 0.3 mm. The correlation between the nn-TRE and gt-TRE matched or exceeded the best performance from manual optimization of the cost function for each of the four contours examined, shown in Table III.

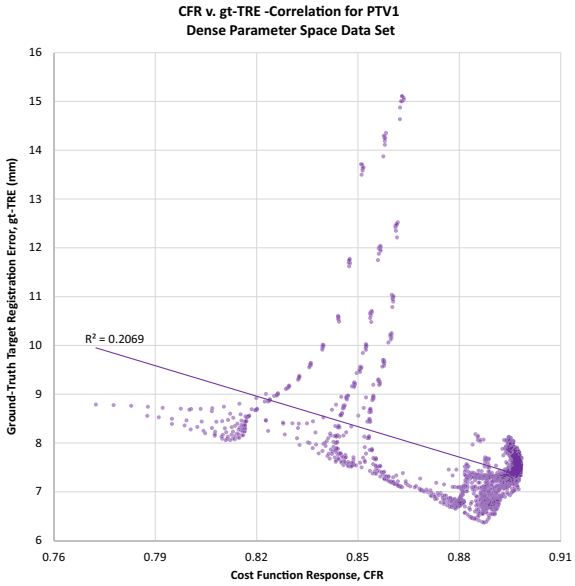

FIG. 5. Plotting the cost function response against the target registration error for the PTV1 contour. The data plotted here corresponds to the data plotted in figure 8.1. [Color figure can be viewed at wileyonlinelibrary.com]

Figure 8 plots the nn-TRE with respect to the gt-TRE, in comparison to Fig. 4, showing much better correspondence with a tight grouping along the linear trend-line. Results show the neural network can accurately predict the TRE for a large range of registration parameter combinations and the resultant range in registration quality. This shows potential for automated registration optimization, with the level of specificity (patient, site) determined by the annotated training data. These experiments indicated the neural network could be trained to identify the best set of registration parameters for a single deformation.

3.B.2. Training on the multi-pose anatomy data set

The next experiment was developed to test whether the neural network could be trained to predict the registration error for a variety of different anatomies that could be seen from day-to-day in the clinic. The multi-pose anatomy data set consisted of 45 different postures, and applied six levels of tumor regression at each posture, providing a good representation of possible anatomies that could be seen from day to day in the clinic. Registrations were performed for five

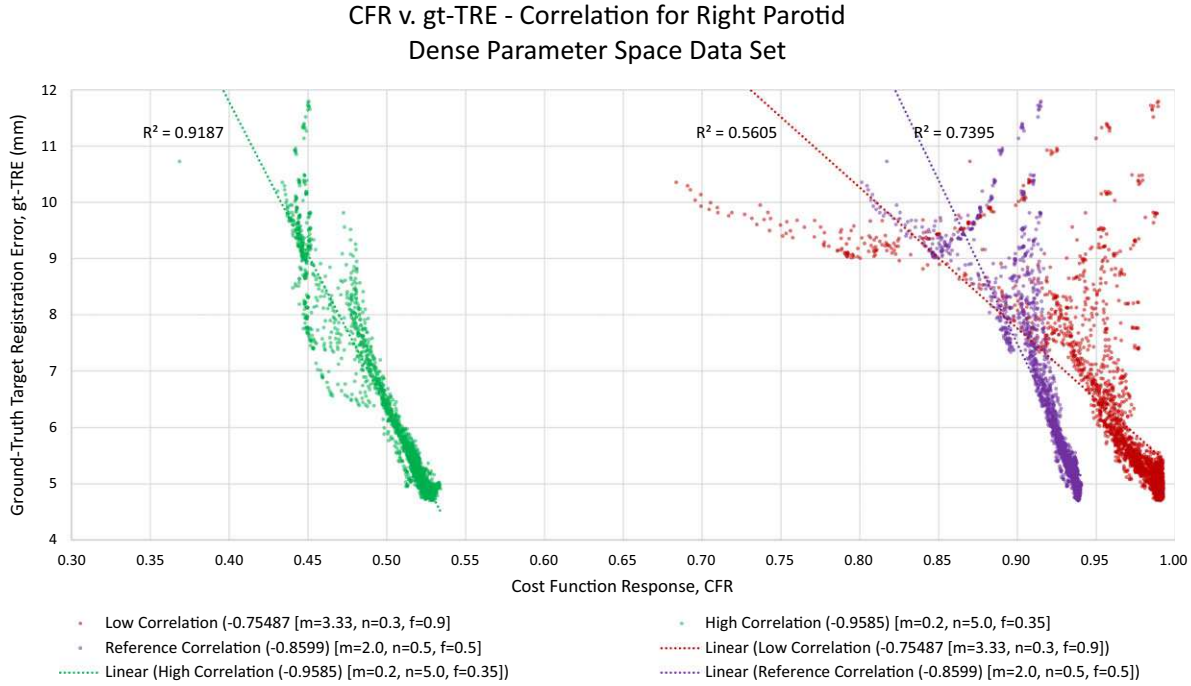

FIG. 6. Correlation between target registration error (TRE) and cost function response (CFR) for three sets of cost function variables (CFV) using the full registration parameter space data, illustrating the high variability of response observed by adjusting the CFV. [Color figure can be viewed at wileyonlinelibrary.com]

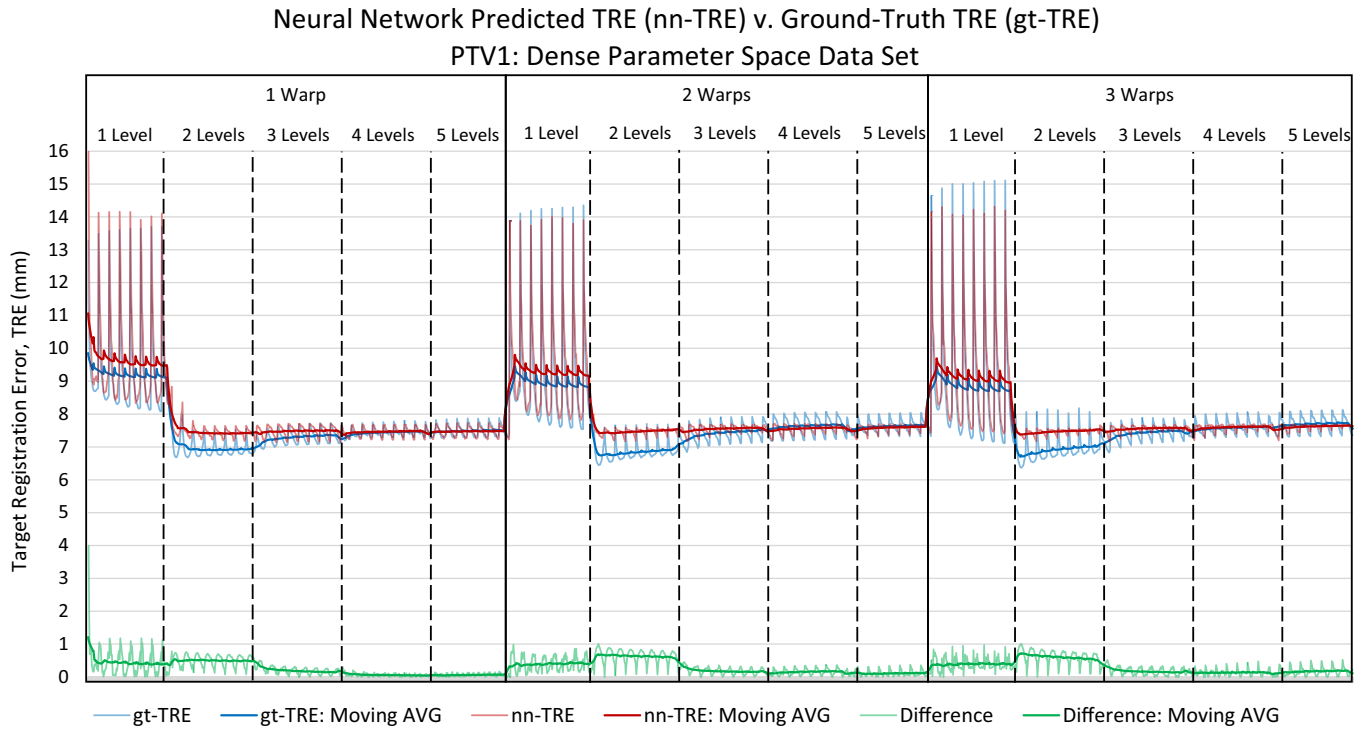

FIG. 7. Comparison of gt-TRE and nn-TRE over the entire dense parameter space data set for the PTV1 contour, with moving averages for a window size of 20 samples. The neural network was able to predict the TRE to within 1 mm after being trained on 25% of the annotated ground truth data. The difference in mm is also plotted with its moving average. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE III. Correlation with ground-truth TRE for cost function response before and after optimizing the cost function variables, and for the neural network predicted TRE, trained on 25% of the Dense Parameter Space data set.

	CFR v. gt-TRE		
	Reference CFV	Best CFV	nn-TRE v. gt-TRE
PTV1	−0.467	−0.649	0.950
Left parotid	−0.921	−0.952	0.988
Right parotid	−0.860	−0.958	0.952

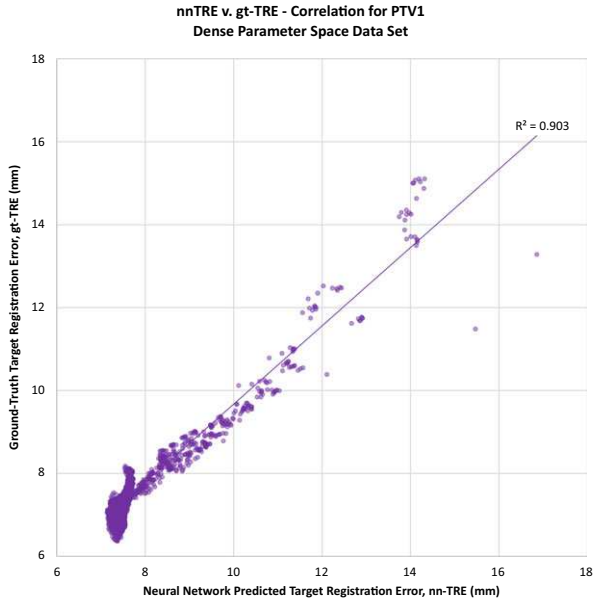

FIG. 8. Plotting the nn-TRE against gt-TRE of the PTV1 contour for the dense parameter space data set. [Color figure can be viewed at wileyonlinelibrary.com]

different smoothing values ranging from 50 to 1000 for each posture. The smoothing variable dictated the scope of local continuity for the deformation vector field. The other registration parameters were set to constant values that are reasonable for clinical registrations. Therefore, the multi-pose anatomy data set consisted of clinically realistic day-to-day anatomies, with relatively small deformations, where DIR performance was expected to be good. For each pose, only subtle differences were expected between registrations based on the different smoothing parameters.

The architecture of the neural network remained the same for both experiments. The network was trained on 25% of the multi-pose anatomy data set, and achieved over 95% accuracy on the test data. The results for the PTV1 contour are shown in Fig. 9. It is apparent from the figure that the registrations as a whole were much better than the registrations in the dense parameter space data set, with the moving average of the gt-TRE ranging from 0.6 and 1.3 mm. However, there was still a large amount of variation between registrations, and the neural network was able to accurately reproduce the high frequency fluctuations with an average error of less than 0.1 mm. The only instance of significant deviation between

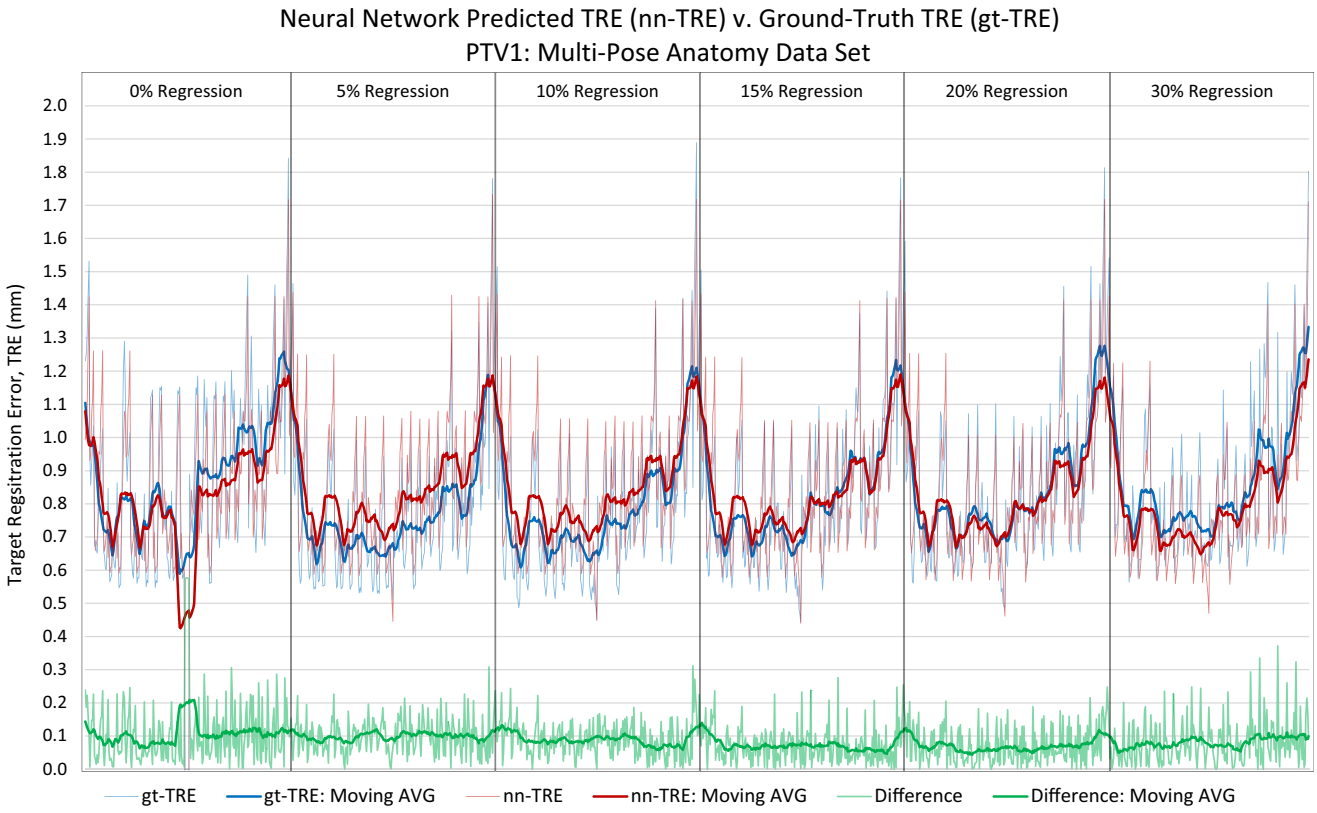

FIG. 9. Comparison of gt-TRE and nn-TRE over the entire multi-pose anatomy data set for the PTV1 contour, with moving averages for a window size of 20 samples. The neural network maintained an average error of approximately 0.05 mm after being trained on 25% of the annotated ground truth data. The difference in mm is also plotted with its moving average. The segmentation of regression level are shown on the figure. Within each regression segment are 45 different postures, each of which was registered five times with different smoothing parameters. [Color figure can be viewed at wileyonlinelibrary.com]

the nn-TRE and the gt-TRE was for the case of no deformation seen in the middle of the 0% regression block. The correlation between the nn-TRE and gt-TRE for the multi-pose anatomy data set was 0.889 for the PTV1, 0.95 for the right parotid, and 0.945 for the left parotid.

This experiment was repeated for nine additional patients with comparable accuracy. A multi-pose anatomy data set was generated for each of the ten patients, which was then used to train a patient-specific neural network. Table IV displays the overall accuracy of the nn-TRE for each of the ten patients, as well as the correlation between the nn-TRE and gt-TRE for the parotid glands and PTV of each patient. The average accuracy of all ten patients was greater than 90%, with correlations over 0.95 for each parotid and over 0.9 for the PTV. These results support the hypothesis that a neural network can reliably infer TRE from only image similarity information for patient-specific scenarios, when properly trained using annotated data. The considerations for size and scope of the training data, as well as the potential avenues for incorporating a neural network are addressed in the discussion.

4. DISCUSSION

The work presented in this paper discusses the feasibility of an automated, quantitative estimation of DIR accuracy for clinical radiotherapy using synthetic data generated from a biomechanical model. Conventionally, TRE calculations have used manually placed landmarks as the gold standard for DIR accuracy measurements, but this landmark-based approach is inapplicable for online ART because of its inherent time complexity and inter-observer variations. Image similarity metrics can be calculated in real-time but are limited by their lack of correspondence to physical units. We hypothesized that a nonlinear relationship between the ISM expectations and the TRE can be modeled using a neural network based approach.

4.A. Summary of contributions

Through supervised training of deep neural networks on large patient-specific, model-generated data sets, we inferred an estimated DIR TRE from computed ISMs. Once trained, the network requires only image similarity information in order to provide a robust, quantified confidence measure of DIR performance in near real-time. For reasonable registrations applied to a variety of possible daily deformations, the network achieved greater than 95% accuracy when comparing its inferred TRE to ground-truth TRE. The key components of this method are (a) the development of the ISM expectation equations, (b) the deep learning framework that translates the ISM values to physical distance metric, and (c)

TABLE IV. Summarized results of neural network performance trained on multi-pose anatomy data sets generated from 10 patients.

	Accuracy		Correlation		
Patient	%	mm	Left parotid	Right parotid	PTV
1	94.76	0.022	0.9452	0.9399	0.8924
2	82.43	0.249	0.9266	0.9214	0.9017
3	93.66	0.032	0.9654	0.9588	0.8986
4	94.28	0.022	0.9500	0.9571	0.8421
5	95.38	0.019	0.9862	0.9745	0.9740
6	95.25	0.021	0.9847	0.9504	0.9566
7	86.50	0.227	0.9560	0.9486	0.9526
8	86.91	0.134	0.9442	0.9401	0.9472
9	92.55	0.054	0.9568	0.9562	0.9015
10	91.61	0.057	0.9751	0.9722	0.7474
Average	91.33	0.084	0.9590	0.9519	0.9014

the generation of clinically realistic ground-truth deformation suites for training patient-specific neural networks.

4.B. Transitioning from proof-of-concept to clinical implementation

While the results presented in this study are promising and highlight the versatility and potential of a neural network based approach to DIR performance assessment, there remains much work to do before clinical implementation. The biomechanically induced deformations of a patient are generated from a single source CT, so the resultant synthetic CTs have similar image characteristics. Assessing a registration of clinical data using a DNN trained only on model-generated deformation data would likely result in a loss of accuracy in the nn-TRE. As discussed below in Section C, network performance is completely dependent on the annotated training data. Therefore, to implement such a method clinically, the training data would need to be expanded. To what extent is a difficult question that requires systematic study. Section E discusses future work and how we intend to move towards clinical implementation.

4.C. Dependence on annotated training data

While neural networks appear to have great potential for fast, quantitative estimates of DIR performance, they are completely dependent on the accuracy and reliability of their annotated training data. The role of biomechanical head and neck models as a key source of training data is critical in enabling such deep neural networks to be instantiated. Using a model, it would be feasible to generate a suite of postures like the multi-pose anatomy data set, for each patient in the clinic. Generating the suite of different postures, and running the registrations to produce the annotated TRE data took approximately one full day using a GPU-based biomechanical model and a fast optical flow deformable image registration algorithm. Network training over 1000 epochs took only a couple minutes. Once trained, the network inferred an estimated TRE almost instantaneously, once the image similarity had been calculated.

4.D. Flexibility of the neural network approach

While the network training relies heavily on the model-generated annotated data, it should be noted that this neural network approach is independent of the biomechanical model used in this study. Any model with the capability to reproduce clinically observed deformations could provide annotated training data, and improvements to biomechanical modeling in general should easily be incorporated in the future.

Similarly, the normalized mutual information image similarity metric could be replaced or combined with any number of other similarity metrics. Although the accuracy and effectiveness of those similarity metrics is yet undetermined.

Lastly, the deformable image registration algorithm is also interchangeable. As evidenced by the work of Kirby et al.,[29] there is a need in radiation therapy for task or application specific registration algorithms. This neural network method could be employed to test and compare DIR algorithms for patient-specific registration tasks, and help ensure the proper algorithm was applied.

4.E. Future work

Future work would focus on using ISMs directly for DIR to enable the DIR's parameter space optimization. While the role of such a parameter space optimization has been demonstrated using TRE [50], the formulation presented in this paper enables the usage of the optimization for scenarios where the TREs are not typically available. This manuscript focused on quantifying the expected error in the deformable image registration, but a similar methodology could be applied for registration optimization. This again delves into the pre-determination of how and when the neural network should be employed. It may be possible to train a network to choose the best combination of registration parameters based on image similarity analysis of the source and target images. Once registered, a second network would give a quantified confidence of the registration performance. Alternatively, the network predicted TRE could be incorporated into a feedback loop with the registration algorithm for task or site-specific optimization.

Contour specific results were calculated in this manuscript by analyzing the sub-volumes encompassing structures of interest. This provided more information than a single measure of similarity between the entire 3D data volumes, and limited focus to the areas of greatest importance. Future work will investigate the calculation of a volumetric image similarity using a moving window throughout the entire 3D image set, similar to the application of a convolution filter. Greater weight can still be given to contours of interest, while delivering more detailed information. Coupled with a neural network, this could produce a volumetric measure of the DIR

confidence, able to be viewed as a heat map and identify problem areas or to adapt the registration to the clinical task. Additionally, this could lead to adaptive registration parameters such as heterogeneous smoothing values, which will be investigated separately.

Within the context of further developing the deep neural networks, future investigation will focus on determining how broad of a scope the deep neural network can have while maintaining accuracy. We focused on analyzing a limited number of critical structures for head-and-neck radiotherapy. Any increase in scope would bring an accompanying requirement for more training data and additional network complexity in the form of more hidden layers or more neurons per layer. Similarly, training individual networks for each structure may improve network results and decrease the amount of training data required. Determining where and how to apply such networks should be an intense area of research, as their applications are wide-ranging and largely unexplored. The inhibiting factor will most likely remain the time and effort required to compile the annotated training data, which further highlights how a fast, versatile, and accurate biomechanical model can be an invaluable resource.

6. CONCLUSION

The machine learning based approach described in this manuscript has the potential to overcome the time and labor hindrances of quantitative DIR error assessment, establish a connection between image similarity to registration error, and provide a fast, automated avenue for clinical DIR validation, which would then facilitate accurate dose accumulation and re-planning, thereby enabling online ART.

CONFLICTS OF INTEREST

The authors have no relevant conflicts of interest to disclose.

[a]Author to whom correspondence should be addressed. Electronic mail. jneylon@mednet.ucla.edu.

REFERENCES

1. Capelle L, Mackenzie M, Field C, Parliament M, Ghosh S, Scrimger R. Adaptive radiotherapy using helical tomotherapy for head-and-neck cancer in definitive and postoperative settings: initial results. *Clin Oncol (R Coll Radiol)*. 2012;24:208–215.
2. Foroudi F, Wong J, Kron T, et al. Online adaptive radiotherapy for muscle-invasive bladder cancer: results of a pilot study. *Int J Radiat Oncol Biol Phys*. 2011;81:765–771.
3. Zeidan O, Huddleston AJ. A comparison of soft-tissue implanted markers and bony anatomy alignments for image-guided treatments of head and neck cancers. *Int J Radiat Oncol Biol Phys*. 2009;76:767–774.
4. Zeidan O, Langen KM. Evaluation of image-guidance protocols in the treatment of head and neck cancers. *Int J Radiat Oncol Biol Phys*. 2007;67:670–677.
5. Lindegaard J, Fokdal L, Nielsen S, Juul-Christensen J, Tanderup K. MRI-guided adaptive radiotherapy in locally advanced cervical cancer from a nordic perspective. *Acta Oncol*. 2013;52:1510–1519.
6. Nijkamp J, Marijnen C, Herk MV, Triest BV, Sonke J. Adaptive radiotherapy for long course neo-adjuvant treatment of rectal cancer. *Radiother Oncol*. 2012;103:353–359.
7. Schwartz D, Garden A, Thomas J, et al. Adaptive radiotherapy for head-and-neck cancer: initial clinical outcomes from a prospective trial. *Int J Radiat Oncol Biol Phys*. 2012;83:986–993.
8. Tuomikoski L, Collan J, Keyrilainen J, Visapaa H, Saarilahti K, Tenhunen M. Adaptive radiotherapy in muscle invasive urinary bladder cancer - an effective method to reduce the irradiated bowel volume. *Radiother Oncol*. 2011;99:61–66.
9. Qi XS, Neylon J, Can S, et al. Feasibility of margin reduction for Level II and III planning target volume in head-and-neck image-guided radiotherapy - dosimetric assessment via a deformable image registration framework. *Curr Cancer Ther Rev*. 2014;10:323–333.
10. Xing L, Siebers J, Keall P. Computational challenges for image-guided radiation therapy: framework and current research. *Semin Radiat Oncol*. 2007;17:245–257.
11. Veiga C, McClelland J, Moinuddin S, et al. Toward adaptive radiotherapy for head and neck patients: feasibility study on using CT-to-CBCT deformable registration for "dose of the day" calculations. *Med Phys*. 2014;41:031703.
12. Sotiras A, Davatzikos C, Paragios N. Deformable medical image registration: a survey. *IEEE Trans Med Imaging*. 2013;32:1153–1190.
13. Crum WR, Hartkens T, Hill DL. Non-rigid image registration: theory and practice. *Br J Radiol*. 2004;77:S140–S153.
14. Yan D, Vicini F, Wong J, Martinez A. Adaptive radiation therapy. *Phys Med Biol*. 1997;42:123–132.
15. Hardcastle N, van Elmpt W, De Ruysscher D, Bzdusek K, Tome WA. Accuracy of deformable image registration for contour propagation in adaptive lung radiotherapy. *Radiat Oncol*. 2013;8:243.
16. Brock KK, Deformable Registration Accuracy C. Results of a multi-institution deformable registration accuracy study (MIDRAS). *Int J Radiat Oncol Biol Phys*. 2010;76:583–596.
17. Fabri D, Zambrano V, Bhatia A, et al. A quantitative comparison of the performance of three deformable registration algorithms in radiotherapy. *Z Med Phys*. 2013;23:279–290.
18. Hoffmann C, Krause S, Stoiber EM, et al. Accuracy quantification of a deformable image registration tool applied in a clinical setting. *J Appl Clin Med Phys*. 2014;15:4564.
19. Mencarelli A, van Kranen SR, Hamming-Vrieze O, et al. Deformable image registration for adaptive radiation therapy of head and neck cancer: accuracy and precision in the presence of tumor changes. *Int J Radiat Oncol Biol Phys*. 2014;90:680–687.
20. Varadhan R, Karangelis G, Krishnan K, Hui S. A framework for deformable image registration validation in radiotherapy clinical applications. *J Appl Clin Med Phys*. 2013;14:4066.
21. Tilly D, Tilly N, Ahnesjo A. Dose mapping sensitivity to deformable registration uncertainties in fractionated radiotherapy - applied to prostate proton treatments. *BMC Med Phys*. 2013;13:2.
22. Fitzpatrick JM, West JB. The distribution of target registration error in rigid-body point-based registration. *IEEE Trans Med Imaging*. 2001;20:917–927.
23. Woods RP, Grafton ST, Holmes CJ, Cherry SR, Mazziotta JC. Automated image registration: I. General methods and intrasubject, intramodality validation. *J Comput Assist Tomogr*. 1998;22:139–152.
24. Woods RP, Grafton ST, Watson JD, Sicotte NL, Mazziotta JC. Automated image registration: II. Intersubject validation of linear and nonlinear models. *J Comput Assist Tomogr*. 1998;22:153–165.
25. van Rijssel MJ, Dahele M, Verbakel WF, Rosario TS. A critical approach to the clinical use of deformable image registration software. In response to Meijneke et al.. *Radiother Oncol*. 2014;112:447–448.
26. Crum WR, Hill DL, Hawkes DJ. Information theoretic similarity measures in non-rigid registration. *Inf Process Med Imaging*. 2003;18:378–387.
27. Rohlfing T. Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable. *IEEE Trans Med Imaging*. 2012;31:153–163.
28. Stanley N, Glide-Hurst C, Kim J, et al. Using patient-specific phantoms to evaluate deformable image registration algorithms for adaptive radiation therapy. *J Appl Clin Med Phys*. 2013;14:4363.
29. Kirby N, Chuang C, Ueda U, Pouliot J. The need for application-based adaptation of deformable image registration. *Med Phys*. 2013;40:011702.

30. Nie K, Chuang C, Kirby N, Braunstein S, Pouliot J. Site-specific deformable imaging registration algorithm selection using patient-based simulated deformations. *Med Phys*. 2013;40:041911.
31. Kirby N, Chen J, Kim H, Morin O, Nie K, Pouliot J. An automated deformable image registration evaluation of confidence tool. *Phys Med Biol*. 2016;61:N203–N214.
32. Wu J, Murphy MJ. A neural network based 3D/3D image registration quality evaluator for the head-and-neck patient setup in the absence of a ground truth. *Med Phys*. 2010;37:5756–5764.
33. Wu J, Su Z, Li Z. A neural network-based 2D/3D image registration quality evaluator for pediatric patient setup in external beam radiotherapy. *J Appl Clin Med Phys*. 2016;17:5235.
34. Fonseca P, Mendoza J, Wainer J, et al. Automatic breast density classification using a convolutional neural network architecture search procedure. *SPIE Med Imaging Comput Aid Diagn*. 2015;9414:941428-1–941428-8.
35. Cruz-Roa A, Basavanhally A, Gonzalez F, et al. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. *SPIE Med Imaging Digit Pathol*. 2014;9041:904103-1–904103-15.
36. Wang H, Cruz-Roa A, Basavanhally A, et al. Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features. *J Med Imaging*. 2014;1:034003-1–034003-8.
37. Bar Y, Diamant I, Wolf L, Greenspan H. Deep learning with non-medical training used for chest pathology identification. *SPIE Med Imaging Comput Aid Diagn*. 2015;9414:94140V-1–94140V-7.
38. Roth H, Lu L, Seff A, et al. A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations. Lecture Notes in Computer Science, MICCAI, 8673; 2014: 520–527.
39. Roth H, Yao J, Lu L, Stieger J, Burns J, Summers R. Detection of sclerotic spine metastases via random aggregation of deep convolutional neural network classifications. ArXiv; 2014.
40. Roche A, Malandain G, Pennec X, Avache N. The correlation ratio as a new similarity measure for multimodal image registration. Lecture Notes on Computer Science, 1496(MICCAI'98); 1998: 1115–1124.
41. Wachowiak M, Smolikova R, Peters T. Multiresolution biomedical image registration using generalized information measures. Lecture Notes on Computer Science, 2879(MICCAI'03); 2003: 846–853.
42. Neilsen M. *Neural Networks and Deep Learning*. Determination Press, http://neuralnetworksanddeeplearning.com/; 2015.
43. Kline D, Berardi V. Revisiting squared-error and cross-entropy functions for training neural network classifiers. *Neural Comput Appl*. 2005;14: 310–318.
44. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proceedings of the IEEE; 1998.
45. Rumelhart D, Hinton G, Williams R. Learning internal representations by error propagation. Parallel distributing processing: Exploration in the microstructure of cognition; 1986: 318–362.
46. Bottou L. Large-scale machine learning with stochastic gradient descent. Proceedings of COMPSTAT; 2010: 177–186.
47. Zeiler MD. ADADELTA: an adaptive learning rate method. arXiv:1212.5701v1; 2012.
48. Neylon J, Qi X, Sheng K, et al. A GPU based high-resolution multilevel biomechanical head and neck model for validating deformable image registration. *Med Phys*. 2015;42:232–243.
49. Min Y, Neylon J, Shah A, et al. 4D-CT Lung registration using anatomy-based multi-level multi-resolution optical flow analysis and thin-plate splines. *Int J Comput Assist Radiol Surg*. 2014;9:875–889.
50. Dou TH, Min Y, Neylon J, Thomas D, Kupelian P, Santhanam AP. Fast simulated annealing and adaptive Monte Carlo sampling based parameter optimization for dense optical-flow deformable image registration of 4DCT lung anatomy. Proc. SPIE 9786, Medical Imaging 2016: Image-Guided Procedures, Robotic Interventions, and Modeling, 97860N (March 18, 2016); doi:10.1117/12.2217194.